Perceived Audiovisual Quality of Low-Bitrate Multimedia Content

Stefan Winkler and Christof Faller

Abstract—**This paper studies the quality of multimedia content at very low bitrates. We carried out subjective experiments for assessing audiovisual, audio-only, and video-only quality. We selected content and encoding parameters that are typical of mobile applications. Our focus were the MPEG-4 AVC (a.k.a. H.264) and AAC coding standards. Based on these data, we first analyze the influence of video and audio coding parameters on quality. We investigate the optimal trade-off between bits allocated to audio and to video under global bitrate constraints. Finally, we explore models for the interactions between audio and video in terms of perceived audiovisual quality.**

Index Terms—**Audiovisual quality assessment, multimedia perception, subjective experiments.**

I. Introduction

QUALITY is the main factor driving research in video and audio compression. It is also an important criterion for codec selection. Examples of such quality comparisons of state-of-the-art video and audio codecs can be found in [1]–[3] for various applications.

Audio and especially speech quality evaluation have quite a long history. There are several subjective testing standards [4], [5]. Additionally, speech and audio quality metrics have been standardized in the form of perceptual evaluation of speech quality (PESQ) [6] and perceived audio quality (PEAQ) [7], respectively.

Video quality evaluation [8] has also become a well-established research area. Standards for subjective assessment [9], [10] have been around for many years, and the International Telecommunication Union (ITU) recently recommended several full-reference quality metrics for TV applications [11], [12] based on the work of the Video Quality Experts Group (VQEG).

Audiovisual (AV) quality, however, is a relatively unexplored topic. An overview of various types of interaction between these two modalities is given in [13]. Previous studies have investigated teleconferencing based on H.261 [14], content with analog distortions [15], MPEG-2 broadcasting [16], and video telephony [17]. There has also been a significant amount of work on audio-video synchronization requirements a.k.a. lip sync [18].

This paper focuses on very low bitrates achievable with today's codecs for mobile applications such as MMS or video streaming. They are characterized by a specific set of requirements that include low bitrates, small frame sizes, and low frame rates. Furthermore, the video is viewed at short distance on a small LCD screen with a progressive display.

For our experiments, we selected source material covering a representative set of content. The source clips were encoded with codecs well-suited for 3G mobile applications, namely MPEG-4 AVC [19], also known as H.264 [20], traditional MPEG-4 [21] and H.263 [22] for video as well as MPEG-4 AAC [23] for audio. Bitrates ranged from 24 to 48 kb/s for video and from 8 to 32 kb/s for audio, based on the fact that a 64 kb/s link is commonly used for circuit-switched video delivery. Bitrates achievable over various other mobile delivery options are similar [24].

Subjective ratings were obtained for the resulting test clips for audio-only, video-only, and audiovisual presentations using the absolute category rating (ACR) methodology defined by ITU-T Recommendations P.910 [10] and P.911 [25]. Based on the quality ratings obtained in these tests, this work addresses the following questions [26], [27]:

- What are the effects of the video codec and the frame rate on video quality?
- What are the effects of the number of audio channels (mono or two-channel stereo) and the sampling rate on audio quality?
- What is the optimal tradeoff between audio and video bit budget to achieve the maximum overall quality?
- How do perceived audio and video quality relate and combine to perceived audiovisual quality?

The paper is organized as follows. Section II introduces the source video and audio content, the simulation environment, and the test conditions used to generate the test material. Section III describes the subjective assessment method, the viewing setup and the presentation structure. Section IV discusses the subjective data, the influence of video and audio coding parameters on perceived quality, as well as the optimal audio-video bit budget allocation. The modeling of the overall audiovisual quality as a function of audio and video quality is the topic of Section V.

II. Test Material

A. Source Clips

The content of the source clips and the range of coding complexity was chosen to be representative of a typical scenario for

Manuscript received March 20, 2005; revised November 21, 2005. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yong Rui.

S. Winkler is with Genista Corporation, Singapore 068 641, and also with the National University of Singapore, Singapore (e-mail: stefan.winkler@genista.com).

C. Faller is with the Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland (e-mail: christof.faller@epfl.ch).

Digital Object Identifier 10.1109/TMM.2006.879871

TABLE I
VIDEO AND AUDIO CONTENT OF SOURCE CLIPS

Clip	Name	Video	Audio	Duration
A	Buildings	slow horizontal pan across a city skyline, followed by a vertical pan up a building facade	orchestral background music	7.48 sec.
B	Conversation	camera switching between head-and-shoulders shots of a woman and a man talking	male and female voices	8.36 sec.
C	Football	American football scene from VQEG [28]; high motion	crowd cheering and chanting; female commentator	7.60 sec.
D	Music video	music video clip; high motion	rock music with vocals	8.08 sec.
E	Trailer 1	action movie trailer; scene cuts and high motion	theme music and voice-over	8.84 sec.
F	Trailer 2	romance movie trailer with credits; scene cuts	theme music and voice-over	8.08 sec.

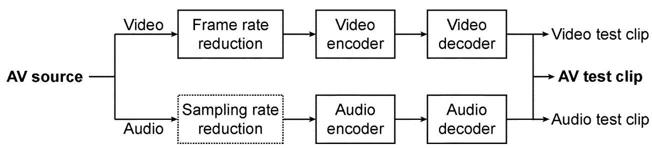

Fig. 1. Encoding setup. Video and audio are processed separately and joined only after decoding.

TABLE II
VIDEO TEST CONDITIONS

Condition	Codec	Frame rate	Bitrate
1	H.264	8 fps	24 kb/s
2	H.264	8 fps	32 kb/s
3	H.264	8 fps	40 kb/s
4	H.264	8 fps	48 kb/s
5	H.263	8 fps	48 kb/s
6	MPEG-4	8 fps	48 kb/s
7	H.264	15 fps	24 kb/s
8	H.264	15 fps	32 kb/s
9	H.264	15 fps	40 kb/s
10	H.264	15 fps	48 kb/s

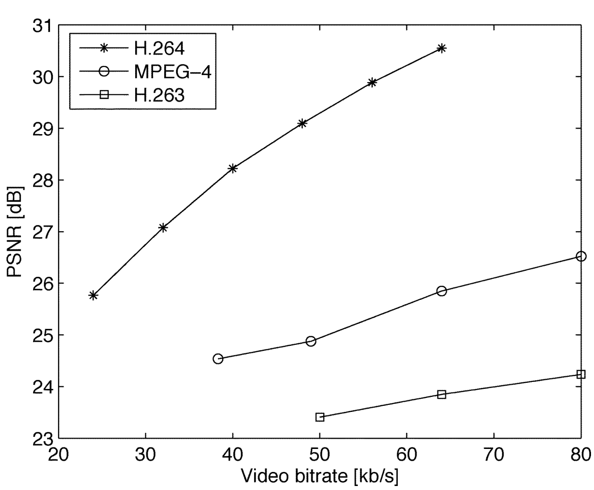

Fig. 2. Peak signal-to-noise ratio (PSNR) of clip D ("Music video") as a function of bitrate for JM H.264 (stars), QuickTime MPEG-4 (circles), and H.263 (squares). The QuickTime encoders did not achieve the target bitrates at the low end of the range.

watching video on a mobile device. The source material comprises six clips of about 8 s each. The video and audio content of these clips is summarized in Table I. All sources except clip C are used with their original audio; an appropriate sound track was added to clip C.

TABLE III
AUDIO TEST CONDITIONS

Condition	Channels	Sampling rate	Bitrate
1	mono	8 kHz	8 kb/s
2	mono	16 kHz	16 kb/s
3	mono	22 kHz	24 kb/s
4	mono	32 kHz	32 kb/s
5	mono	22 kHz	32 kb/s
6	stereo	22 kHz	32 kb/s
7	stereo	16 kHz	32 kb/s

The video source material was originally in TV format; for our tests we de-interlaced and downsampled it to QCIF frame size (176×144). The audio source material was 16-bit PCM stereo sampled at 48 kHz.

B. Test Conditions

Codec selection was principally determined by the 3GPP[1] file format as defined in [29]. It is of particular interest for packet-switched video streaming in 3G networks.

The encoding setup is shown in Fig. 1. Before encoding, the video frame rate of the source clips was reduced to 8 or 15 fps using VirtualDub.[2] The audio sampling rate reduction was carried out internally by the encoder.

The video conditions are listed in Table II. We chose the MPEG-4 AVC/H.264 [19], [20] coding standard (baseline profile), as well as traditional MPEG-4 part 2 [21] and H.263 [22]. The JM reference software[3] version 8.5 was used for H.264 encoding; Apple QuickTime Pro version 6.5 was used for H.263 and MPEG-4 encoding.

The reason for using H.264 in almost all test conditions is that the QuickTime encoders (especially H.263) did not produce substantial quality variations within the bitrate range of interest; viewers were unable to discern the quality of the different test clips. Furthermore, they did not achieve the target bitrates at the low end of the range. This is demonstrated in Fig. 2. The H.264 JM reference encoder does not have these problems.

The audio conditions are listed in Table III. We chose the MPEG-4 AAC-LC (low complexity) coding standard [23]. QuickTime Pro version 6.5 was again used for encoding, with the "recommended" sampling rate for each target bitrate.

Video conditions 1–4 from Table II were then combined with audio conditions 1–4 from Table III for a total of eight audio-visual test conditions as listed in Table IV. The corresponding

[1] 3rd Generation Partnership Project, see http://www.3gpp.org/.

[2] VirtualDub is available at http://www.virtualdub.org/

[3] The JM reference software is available at http://bs.hhi.de/-suehring/tml/

TABLE IV
AUDIOVISUAL TEST CONDITIONS (Video + Audio)

Condition	Total	Video	Audio
1+2	40 kb/s	24 kb/s	16 kb/s
2+1	40 kb/s	32 kb/s	8 kb/s
1+4	56 kb/s	24 kb/s	32 kb/s
2+3	56 kb/s	32 kb/s	24 kb/s
3+2	56 kb/s	40 kb/s	16 kb/s
4+1	56 kb/s	48 kb/s	8 kb/s
3+4	72 kb/s	40 kb/s	32 kb/s
4+3	72 kb/s	48 kb/s	24 kb/s

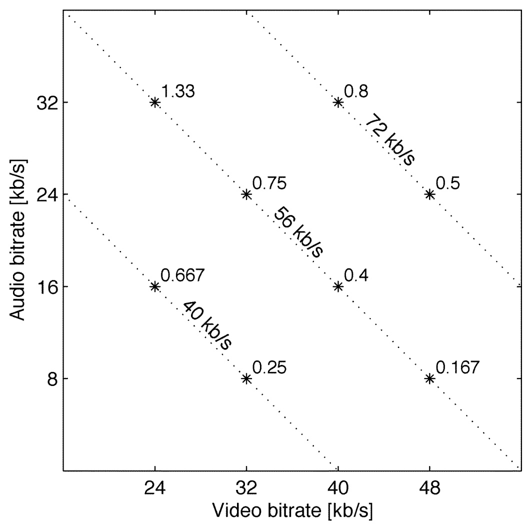

Fig. 3. Audiovisual test conditions. Stars denote the video and audio bitrate combinations used in the test. Diagonal dotted lines connect points with the same total data rate. Every point is labeled with its A/V bitrate ratio.

sampling of the AV bitrate space is illustrated in Fig. 3. Of particular interest is a total data rate of 56 kb/s, which may be transmitted over a typical 64 kb/s circuit-switched connection (including bitstream packetization overhead).

III. SUBJECTIVE ASSESSMENT

A. Assessment Method

The experimental setup follows ITU-T Recommendations [10], [25]. We use ACR, a very efficient testing methodology, where the test clips are viewed one at a time and rated independently on a discrete 11-level scale from “bad” (0) to “excellent” (10). The ratings for each test clip are then averaged over all subjects to obtain a mean opinion score (MOS).

Our initial plan was to use hidden reference removal as proposed by some studies [30] as well as upcoming VQEG evaluations for single stimulus tests. Hidden reference implies that the subjects are not aware of the fact that the original uncompressed clips are included in the test. The “removal” of the hidden reference is done in the analysis by subtracting each subject’s score for the reference from the corresponding test clips. However, we found the quality difference between reference and compressed clips to be so large that we decided against including the reference clips in the set evaluated by the subjects.

B. Subjects

Twenty–four subjects (six female, 18 male) participated in the test. Their age ranged from 25 to 36 years. Four subjects were familiar with image processing, one was familiar with audio processing. All subjects had normal or corrected vision and normal hearing.

C. Setup

The tests were conducted in a dark and sound-insulated room. The monitor used in the subjective assessments was a 17″ LCD screen (Dell 1703FP) at its native resolution of 1280 $\times$ 1024 pixels. The video clips were displayed at their original size (QCIF) in the center of the screen, surrounded by a uniform gray background. The viewing distance was not fixed. For our test material, we found subjects to be most comfortable at a viewing distance of around 30–40 cm, which corresponds to about 8–10 times the height of the video picture in our setup.

For the audio playback, an external D/A converter (Emagic EMI A26) was used. High-quality headphones (Sennheiser HD 600) were directly connected to the D/A converter.

Genista’s *QualiView* software was used for the playback of the test clips. It reads the decoded clips (both video and audio) stored in uncompressed AVI format and plays them out (audio or video can be switched on and off separately). After each clip, the voting dialog shown in Fig. 4 is presented on the screen, and the rating entered by the subject is recorded.

D. Presentation Structure

Written instructions were given to the subjects at the beginning of the test session, explaining the three-tiered structure of the session as well as the voting task and dialog. An instructor was present to answer questions.

A short training session preceded the actual test; it comprised three audiovisual clips demonstrating the extremes of expected audio and video quality ranges.[4] The subjects were allowed to adjust the viewing distance and the headphone volume during the training session.

The actual test took about 30 min and consisted of three consecutive parts, which are listed in Table V. The audiovisual part came first, followed by the audio-only and video-only presentations. This order was chosen for a number of reasons: We wanted to minimize fatigue during the AV part, which we considered the most important. Also, the AV test clips comprised only a subset of the audio-only and video-only test conditions; consequently, the later parts of the test still contained clips that subjects had not seen or heard before. From a more practical point of view, this order allowed subjects to remove the headphones after the second part.

The subjects were asked to rate the quality of the presentation in each of the three parts. Subjects were allowed to take a short break between the different parts and continue when they were ready. The order of the clips within each part was randomized individually for each subject.

[4] The training clips were taken from the test set of audio and video clips, but their AV combinations did not occur in the test. Specifically we used the following source-video/audio combinations: F-6/3, D-1/1, A-4/6.

TABLE V
SUBDIVISION OF SUBJECTIVE TEST SESSION IN THREE PARTS

#	Part	Result	Conditions	Clips	Comment
1	Audiovisual quality (AVQ)	MOS_{AV}	Table IV	48	
2	Audio-only quality (AQ)	MOS_{A}	Table III	42	blank (gray) screen
3	Video-only quality (VQ)	MOS_{V}	Table II	60	muted audio

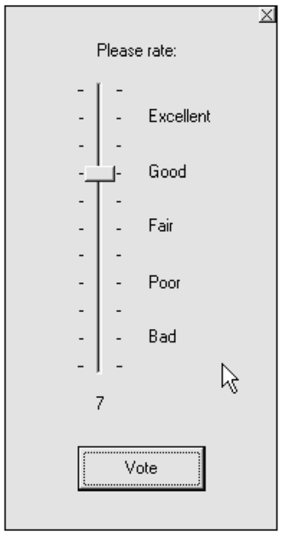

Fig. 4. ACR voting dialog.

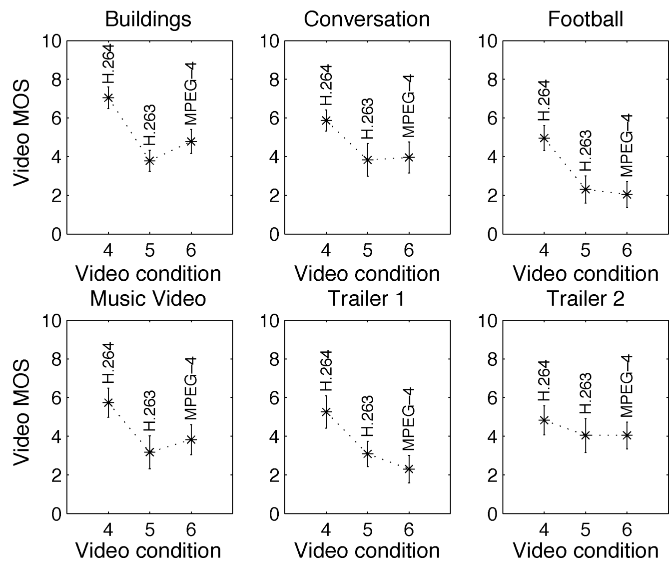

Fig. 5. Video MOS comparison for different codecs. Error bars indicate the 95%-confidence intervals.

IV. Test Results

A. Subjective Data Analysis

An analysis of the raw subjective data reveals that the video quality (VQ) variation with bitrate is not very large, but the source clip has a big influence on perceived quality. The opposite is true for audio quality (AQ), where a big difference between condition 1 (8 kb/s) and the others is observed. Subjects hesitated to use the entire range of the ACR scale, especially for the VQ and audiovisual quality (AVQ) parts of the test. MOS values below two and above eight are rare.

The sizes of the 95%-confidence intervals of the subjective data lie between 0.4 and 0.9 for all three tests. This is comparable to other tests and indicates a good agreement between subjects, despite the use of an absolute rating methodology.

B. Video Codecs

To compare the performance of the three codecs in terms of perceived quality, we now look at video test conditions 4–6 from Table II. The VQ MOS values plotted in Fig. 5 show that H.264 clearly outperforms the two other codecs. The only exception is perhaps "Trailer 2", in which H.264 has a hard time coping with the scene cuts. No clear winner can be determined between H.263 and MPEG-4.

Using paired t-tests, we tested the null hypothesis of equal means for each of the three possible codec pairs separately. The resulting p-values are shown in Table VI. They confirm that the QuickTime H.263 and MPEG-4 codecs are not significantly different in visual quality, while JM H.264 is clearly better than both. However, we note that the H.264 JM reference encoder implementation is almost 100 times slower than the two QuickTime encoders.

TABLE VI
T-TEST RESULTS FOR DIFFERENT CODEC PAIRS

Codecs	p-value
H.263 vs. MPEG-4	0.442
H.263 vs. H.264	0
H.264 vs. MPEG-4	0

C. Video Frame Rate

Video test conditions 1–4 and 7–10 from Table II differ only in frame rate (8 fps and 15 fps, respectively). The VQ MOS and 95%-confidence intervals for these conditions are shown in Fig. 6. In most cases, the perceived video quality is markedly better for 8 fps than for 15 fps at the same bitrate. The difference is least pronounced for the low-motion "Conversation" clip, but interestingly also for the two high-motion trailers. The latter contain the most scene cuts, to the extent that the effectiveness of the motion prediction is affected by lowering the frame rate.

Again we carried out paired t-tests of the null hypothesis that 8 fps and 15 fps come from equal means at each bitrate. The resulting p-values, shown in Table VII, lead to a clear rejection of the null hypotheses, thus indicating that a frame rate of 8 fps results in significantly higher video quality than 15 fps at a given bitrate. This confirms previous studies [31].

D. Audio Channels and Sampling Rate

We now study the impact of various audio coding parameters on the perceived audio quality. For this purpose we had included four audio test conditions with the same bitrate (32 kb/s) but varying parameters in the test (conditions 4–7 in Table III). We also include condition 3 in this analysis, as it only differs from

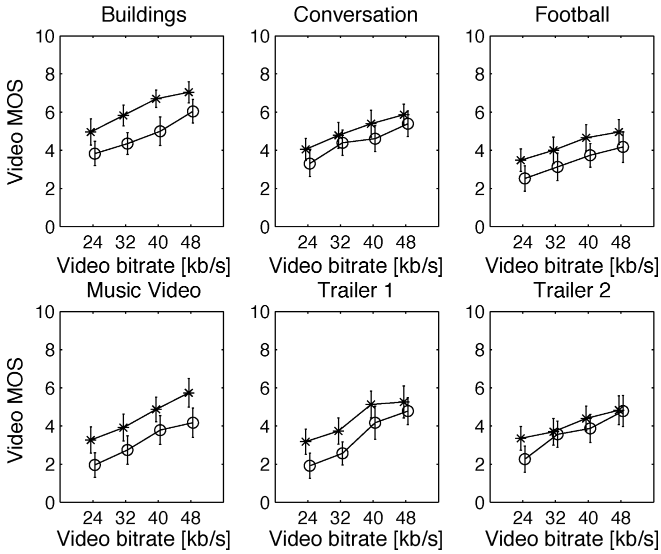

Fig. 6. Video MOS as function of bitrate at 8 fps (stars) and 15 fps (circles). Error bars indicate the 95%-confidence intervals.

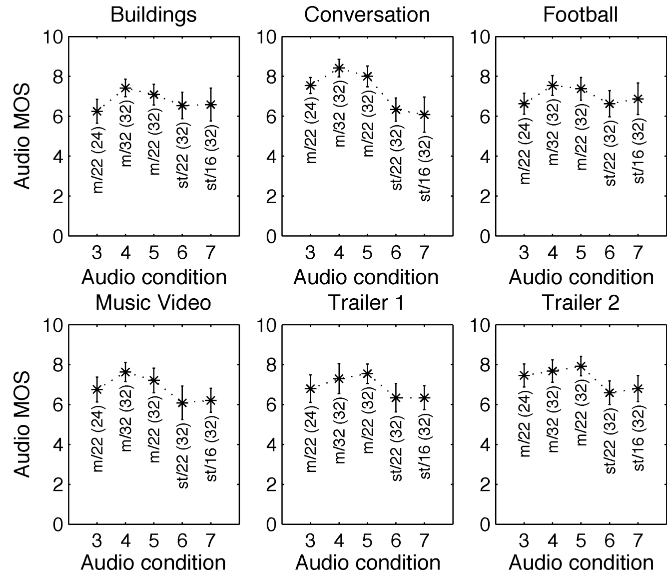

Fig. 7. Audio MOS comparison for mono/stereo, different sampling rates, and two bitrates (see Table III for details). Error bars indicate the 95%-confidence intervals.

TABLE VII
T-TEST RESULTS COMPARING FRAME RATES OF 8 fps AND 15 fps.

Bitrate	p-**value**
24 kb/s	$2.66 \cdot 10^{-14}$
32 kb/s	$1.69 \cdot 10^{-10}$
40 kb/s	$9.72 \cdot 10^{-13}$
48 kb/s	$1.84 \cdot 10^{-6}$

TABLE VIII
T-TEST RESULTS FOR DIFFERENT CODING PARAMETERS

Conditions	p-**value**	**Conditions**	p-**value**
4 vs. 5	0.233	3 vs. 4	$4.39 \cdot 10^{-8}$
4 vs. 6	$1.48 \cdot 10^{-12}$	3 vs. 5	$3.17 \cdot 10^{-3}$
4 vs. 7	$1.25 \cdot 10^{-9}$	3 vs. 6	$2.48 \cdot 10^{-5}$
5 vs. 6	$8.24 \cdot 10^{-11}$	3 vs. 7	$2.18 \cdot 10^{-2}$
5 vs. 7	$8.04 \cdot 10^{-8}$		
6 vs. 7	0.700		

condition 5 in bitrate. The question is how the audio bandwidth (directly related to audio coder sampling rate) and the number of audio channels (mono or two-channel stereo) affect the audio quality.

Fig. 7 shows the AQ MOS for the relevant audio test conditions. For all six clips, the perceived audio quality is higher when mono audio coding is used (conditions 4 and 5) than when stereo audio coding is used (conditions 6 and 7).

We carried out t-tests on all 10 possible condition pairs. The resulting p-values are shown in Table VIII. Condition pairs 4&5 and 6&7 are not significantly different. This implies that changing the audio sampling rate has no measurable effect on quality, regardless of whether mono or stereo is used. However, mono encoding is significantly better than stereo encoding in all four cases. In fact, even 24 kb/s mono is better than 32 kb/s stereo (whereas 32 kb/s mono is always better than 24 kb/s mono).

It is not surprising that the nonparametric transform coder AAC-LC yields better quality for mono considering the low audio bitrates in our test. The audio bandwidth available for two stereo channels is much lower than for a single mono channel when both are coded at the same bitrate. Therefore, the stereo audio appears more distorted, and subjects prefer mono audio with less degradation. The recently standardized high-efficiency AAC (HE-AAC) [32] avoids the issue that at low bitrates only a low audio bandwidth can be afforded for stereo. Using HE-AAC, stereo may be preferred even at these bitrates.

E. Audio-Video Bit Budget Allocation

The AVQ MOS values for the six clips are shown as a function of the audio/video bitrate ratio (cf. Fig. 3) in Fig. 8. Focusing first on 56 kb/s (circles), where we have the most sample points, we note the following. The audio/video bitrate ratio with the highest AVQ depends to a large extent on the specific clip. For five out of the six clips, the optimum ratio is in the center range around 16/40–24/32.

In the visually most complex clips, e.g., "football" and the two trailers, a high relative audio bitrate produces the best overall quality, whereas the less demanding clips ("buildings" and "conversation") benefit from a high video bitrate. This seems counter-intuitive at first, since one would expect complex material to require more bits for the video track. On the other hand, a bitrate increase may result only in a negligible improvement in video quality for such a clip, while an increase by the same amount can significantly improve the audio. This could explain why the bits may in fact be better spent on the audio when the video is very complex.

If the total bitrate budget is reduced to 40 kb/s, the optimum audio/video bitrate ratio decreases, i.e. relatively more bits should be allocated to the video. The opposite trend can be observed when the total bitrate increases to 72 kb/s. In this case, the optimum appears to shift to the right, i.e., a higher relative bitrate for the audio seems favorable. Unfortunately, our test

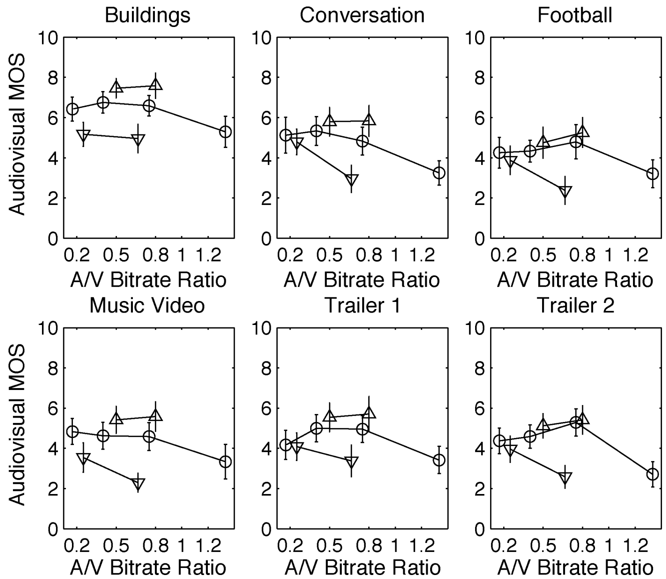

Fig. 8. Audiovisual quality as a function of audio/video bitrate ratio at total bitrates of 56 kb/s (circles), 40 kb/s (downward-pointing triangles) and 72 kb/s (upward-pointing triangles). Refer to Fig. 3 for the exact A/V bitrate ratios of each data point. Error bars indicate the 95%-confidence intervals.

does not include enough data points to draw firm conclusions on this matter.

V. Audio-Video Quality Interactions

A. Principal Component Analysis

To study the influence of AQ, VQ, and the multiplicative interaction term AQ $\cdot$ VQ on AVQ, we carried out a principal component analysis (PCA). Four-dimensional test vectors composed of $\mathrm{MOS_A}$, $\mathrm{MOS_V}$, $\mathrm{MOS_A} \cdot \mathrm{MOS_V}$ and $\mathrm{MOS_{AV}}$ values were constructed. Each vector contains $\mathrm{MOS_A}$ and $\mathrm{MOS_V}$ obtained with the same bitrates as were used for the corresponding $\mathrm{MOS_{AV}}$ item. Prior to the PCA, the mean of the data was removed and the variance was normalized for each dimension.

Fig. 9 shows the eigenvalues corresponding to the four principal components. Since 97% of the variability is contained in the first two principal components, we plot AQ, VQ, AQ $\cdot$ VQ, as well as AVQ vectors relative to the first two principal components in Fig. 10. This plot indicates that neither AQ nor VQ alone determine AVQ; both have a similarly strong influence on AVQ. The multiplicative term AQ $\cdot$ VQ is rather close to AVQ. Note that AQ and VQ are more different from AVQ than in another study [16]. This difference may be due to the small size or the low bitrates of the video clips in our test.

B. Modeling

The PCA described above provides evidence that both AQ and VQ contribute to AVQ. In this section, we further investigate this relationship in terms of modeling and prediction. As other researchers have proposed, $\mathrm{MOS_{AV}}$ can be modeled using

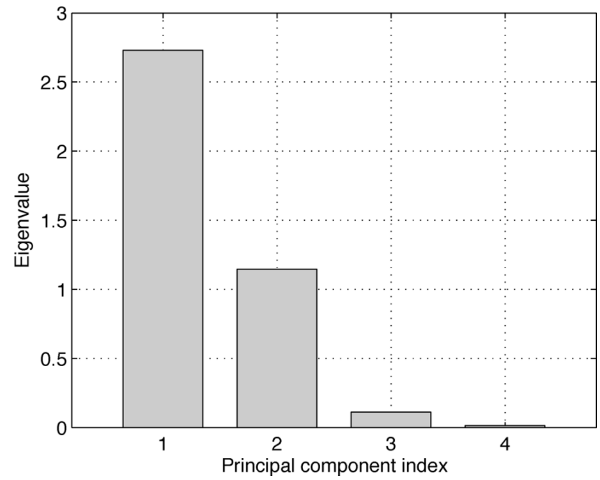

Fig. 9. Eigenvalues of the four principal components.

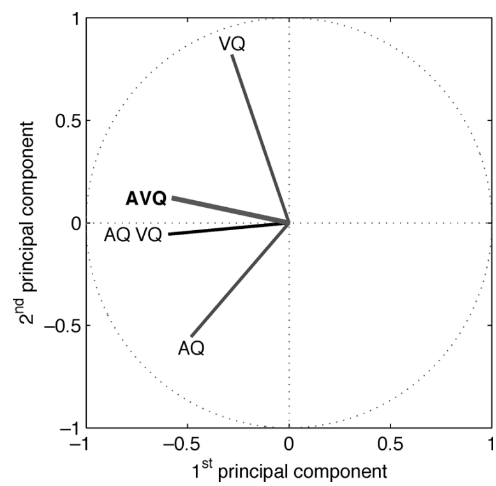

Fig. 10. AQ, VQ, AQ $\cdot$ VQ, as well as AVQ vectors relative to the first two principal components.

$\mathrm{MOS_A}$, $\mathrm{MOS_V}$, and a multiplicative interaction term [15], [17], as follows:

$$\widehat{\mathrm{MOS}}_{\mathrm{AV}} = a_0 + a_1\,\mathrm{MOS_A} + a_2\,\mathrm{MOS_V} + a_3\,\mathrm{MOS_A} \cdot \mathrm{MOS_V}. \tag{1}$$

We apply this model with different numbers of free parameters a_k to our data (a_0 is irrelevant for correlations, but improves the fit in terms of residual). The model accuracy of the various fits is shown in Fig. 11. As expected from the results of the above PCA, good modeling is possible with only the multiplicative term

$$\widehat{\mathrm{MOS}}_{\mathrm{AV}} = 1.98 + 0.103\,\mathrm{MOS_A} \cdot \mathrm{MOS_V} \tag{2}$$

or an additive linear model

$$\widehat{\mathrm{MOS}}_{\mathrm{AV}} = -1.51 + 0.456\,\mathrm{MOS_A} + 0.770\,\mathrm{MOS_V}. \tag{3}$$

The latter provides a somewhat better fit, which is characterized by a correlation of 94% and a root mean square (rms) residual of 0.44. The plane represented by (3) is shown together with

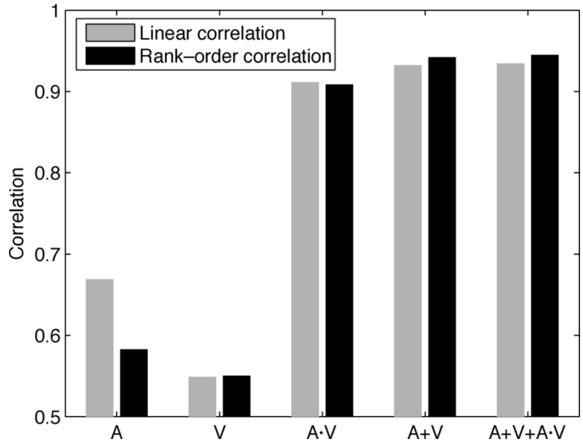

Fig. 11. Correlations of different models for AVQ (left). A: model with $\mathrm{MOS_A}$ only ($a_2 = a_3 = 0$); V: model with $\mathrm{MOS_V}$ only ($a_1 = a_3 = 0$); $\mathrm{A \cdot V}$: multiplicative model from (2); $\mathrm{A + V}$: additive model from (3); $\mathrm{A + V + A \cdot V}$: full model as in (1) with all four parameters.

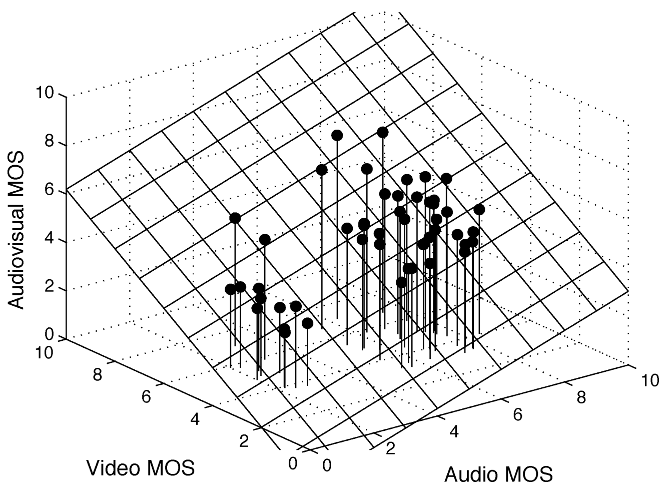

Fig 12. Plane of the AVQ model defined by (3) and $\mathrm{MOS_{AV}}$ values (dots).

the actual $\mathrm{MOS_{AV}}$ values in Fig. 12. It illustrates very well the higher importance attributed to VQ as compared to AQ. There is no improvement when using all four parameters in the fit.

We can compare these fits with other subjective experiments. Much of the existing work has focused on video-conferencing applications (i.e. head-and-shoulders clips), speech, and/or simulated artifacts; the test material used here is quite different in terms of content range and distortions. Despite these significant differences, the multiplicative model from (2) is very similar to previous models [15], [33] in terms of its parameters and goodness of fit. The same can be said about the additive model from (3); the higher weighting of $\mathrm{MOS_V}$ over $\mathrm{MOS_A}$ is confirmed by other studies [15], [17], [33].[5]

VI. Conclusions

We carried out subjective experiments on audio, video, and audiovisual quality using the ACR methodology. Our main interest was the 3GPP format used in mobile video transmission. We focused on MPEG-4 AVC/H.264 and MPEG-4 AAC to encode our test material at very low bitrates (24–48 kb/s for video and 8–32 kb/s for audio).

We investigated the influence of various encoding parameters on audio, video and audiovisual quality under these conditions. The main findings can be summarized as follows.

- The QuickTime Pro encoders for H.263 and MPEG-4 have very similar quality. The H.264 JM reference encoder produces significantly better quality video, but is much slower.
- Encoding at 8 fps produces higher-quality video than 15 fps at the same bitrate.
- Choosing mono instead of stereo produces higher-quality audio with the LC-AAC codec. Changing the sampling rate or even the bitrate has much less effect on the resulting audio quality.
- The optimum audio/video bitrate allocation depends on clip complexity. The more complex the content and the higher the total bitrate budget, the more bits should be allocated to audio. At a total bitrate of 56 kb/s, the optimum is around 32–40 kb/s for video and 16–24 kb/s for audio.

We also found that both AQ and VQ contribute significantly to perceived AVQ. The product of AQ and VQ is an effective model of AVQ, and so is the linear combination of AQ and VQ. Our models confirm the results of previous studies, despite the substantial differences in source material and test conditions. These results can be utilized for the prediction of audiovisual quality by combining VQ and AQ metrics [26].

Future work will include commercial H.264 encoders such as QuickTime version 7 for video and HE-AAC encoders for audio. This will also show if the quality gain of H.264 can be maintained by more efficient codec implementations.

More data points in the AV space are needed for understanding how the optimal audio-video bit budget behaves as a function of overall bitrate. It would also be interesting to see if the observed trend persists at higher bitrates.

Considering the effects of transmission errors encountered in mobile networks is another important aspect, because the resulting artifacts, e.g., audio or video dropouts, can affect perceived quality differently from coding distortions.

Finally, the sensitivity to AV synchronization problems (a.k.a. lip sync) at these low bitrates and their influence of overall perceived quality is also of interest.

Acknowledgment

The authors thank everyone who participated in the subjective experiments; S. Süsstrunk at EPFL's Audiovisual Communications Lab provided some of the subjective testing facilities. They are also grateful to R. Costantini for his comments on statistics and to the anonymous reviewers for their constructive feedback.

References

[1] G. A. Soulodre, T. Grusec, M. Lavoie, and L. Thibault, "Subjective evaluation of state-of-the-art 2-channel audio codecs," *J. Audio Eng. Soc.*, vol. 46, no. 3, pp. 164–177, 1998.

[2] J. Bennett and A. Bock, "In-depth review of advanced coding technologies for low bit rate broadcast applications," in *Proc. Int. Broadcasting Conv.*, Amsterdam, The Netherlands, Sep. 12–16, 2003, pp. 464–472.

[3] F. Kozamernik, P. Sunna, E. Wyckens, and D. I. Pettersen, "Subjective quality of internet video codecs—phase II evaluations using SAMVIQ," *EBU Tech. Rev.*, vol. 301, Jan. 2005.

[5] Hands [17] shows that the content can have an influence on the model parameters, as he finds a stronger weighting of the audio component for video-conferencing material.

[4] General Methods for the Subjective Assessment of Sound Quality, Int. Telecommun. Union, Geneva, Switzerland, ITU-R Rec. BS.1284-1, 2003.
[5] Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems, Int. Telecommun. Union, Geneva, Switzerland, ITU-R Rec. BS.1534-1, 2003.
[6] Perceptual Evaluation of Speech Quality (PESQ), An Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs, Int. Telecommun. Union, Geneva, Switzerland, ITU-T Rec. P.862, 2001.
[7] Method for Objective Measurements of Perceived Audio Quality (PEAQ), Int. Telecommun. Union, Geneva, Switzerland, ITU-R Rec. BS.1387-1, 2001.
[8] S. Winkler, *Digital Video Quality—Vision Models and Metrics*. New York: Wiley, 2005.
[9] Methodology for the Subjective Assessment of the Quality of Television Pictures, Int. Telecommun. Union, Geneva, Switzerland, ITU-R Rec. BT.500-11, 2002.
[10] Subjective Video Quality Assessment Methods for Multimedia Applications. Int. Telecommun. Union, Geneva, Switzerland, ITU-T Rec. P.910, 1999.
[11] Objective Perceptual Video Quality Measurement Techniques for Standard Definition Digital Broadcast Television in the Presence of a Full Reference, Int. Telecommun. Union, Geneva, Switzerland, ITU-R Rec. BT.1683, 2004.
[12] Objective Perceptual Video Quality Measurement Techniques for Digital Cable Television in the Presence of a Full Reference, Int. Telecommun. Union, Geneva, Switzerland, ITU-T Rec. J.144, 2004.
[13] A. Kohlrausch and S. van der Par, "Auditory-visual interaction: from fundamental research in cognitive psychology to (possible) applications," in *Proc. SPIE Human Vision and Electronic Imaging*, San Jose, CA, Jan. 23–29, 1999, vol. 3644, pp. 34–44.
[14] C. Jones and D. J. Atkinson, "Development of opinion-based audiovisual quality models for desktop video-teleconferencing," in *Proc. Int. Workshop on Quality of Service*, Napa, CA, May 18–20, 1998, pp. 196–203.
[15] J. G. Beerends and F. E. de Caluwe, "The influence of video quality on perceived audio quality and vice versa," *J. Audio Eng. Soc.*, vol. 47, no. 5, pp. 355–362, May 1999.
[16] A. Joly, N. Montard, and M. Buttin, "Audio-visual quality and interactions between television audio and video," in *Proc. Int. Symp. Signal Processing and its Applications*, Kuala Lumpur, Malaysia, Aug. 13–16, 2001, pp. 438–441.
[17] D. S. Hands, "A basic multimedia quality model," *IEEE Trans. Multimedia*, vol. 6, no. 6, pp. 806–816, Dec. 2004.
[18] R. Steinmetz, "Human perception of jitter and media synchronization," *IEEE J. Sel. Areas Commun.*, vol. 14, no. 1, pp. 61–72, Jan. 1996.
[19] Coding of Audio-Visual Objects—Part 10: Advanced Video Coding, Int, Org. Standardization, Geneva, Switzerland, ISO/IEC 14496-10, 2004.
[20] Advanced Video Coding for Generic Audiovisual Services, Int. Telecommun. Union, Geneva, Switzerland, ITU-T Rec. H.264, 2003.
[21] Coding of Audio-Visual Objects—Part 2: Visual, Int. Org. Standardization, Geneva, Switzerland, ISO/IEC 14496-2, 2004.
[22] Video Coding for Low Bit Rate Communication, Int. Telecommun. Union, Geneva, Switzerland, ITU-T Rec. H.263, 1998.
[23] Coding of Audio-Visual Objects—Part 3: Audio, Int. Org. Standardization, Geneva, Switzerland, ISO/IEC 14496-3, 2004.
[24] H. Buddendick, A. Weber, and M. Tangemann, "Comparison of data throughput performance in GPRS, EGPRS, and UMTS," in *Proc. World Wireless Congr.*, San Francisco, CA, May 27–30, 2003.
[25] Subjective Audiovisual Quality Assessment Methods for Multimedia Applications. Int. Telecommun. Union, Geneva, Switzerland, ITU-T Rec. P.911, 1998.
[26] S. Winkler and C. Faller, "Audiovisual quality evaluation of low-bitrate video," in *Proc. SPIE Human Vision and Electronic Imaging*, San Jose, CA, Jan. 16–20, 2005, vol. 5666, pp. 139–148.
[27] ——, "Maximizing audiovisual quality at low bitrates," in *Proc. Workshop on Video Processing and Quality Metrics*, Scottsdale, AZ, Jan. 23–25, 2005, invited paper.
[28] VQEG, Final Report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment, Apr. 2000 [Online]. Available: http://www.vqeg.org/
[29] 3GPP Technical Specification 26.244. Transparent End-to-End Packet Switched Streaming Service (PSS); 3GPP File Format (3GP) (Rel. 6). 3rd Generation Partnership Project 2004.
[30] M. Pinson and S. Wolf, "Comparing subjective video quality testing methodologies," in *Proc. SPIE Visual Communications and Image Processing*, Lugano, Switzerland, Jul. 8–11, 2003, vol. 5150, pp. 573–582.
[31] D. Wang, F. Speranza, A. Vincent, T. Martin, and P. Blanchfield, "Towards optimal rate control: a study of the impact of spatial resolution, frame rate, and quantization on subjective video quality and bit rate," in *Proc. SPIE Visual Communications and Image Processing*, Lugano, Switzerland, Jul. 8–11, 2003, vol. 5150, pp. 198–209.
[32] M. Dietz, L. Liljeryd, K. Kjörling, and O. Kunz, "Spectral band replication—a novel approach in audio coding," in *Proc. AES Conv.*, Münich, Germany, May 10–13, 2002.
[33] R. Pastrana-Vidal, C. Colomes, J. Gicquel, and H. Cherifi, "Caractérisation perceptuelle des interactions audiovisuelles: Revue," in *Proc. CORESA Workshop*, Lyon, France, Jan. 16–17, 2003.

Stefan Winkler received the M.Sc. (Dipl.-Ing.) degree in electrical engineering from the University of Technology, Vienna, Austria, in 1996, and the Ph.D. degree in electrical engineering from the Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, in 2000 for work on vision modeling and video quality measurement. He also spent one year at the University of Illinois at Urbana-Champaign as a Fulbright student.

In 2001, he co-founded Genimedia (now Genista Corporation), a company developing perceptual quality metrics for multimedia applications. He later returned to EPFL as a Postdoctoral Fellow and also became assistant professor at the University of Lausanne. He is currently an Assistant Professor at the National University of Singapore and Chief Scientist at Genista Corporation. He has published more than 30 papers on vision modeling and quality assessment and is the author of a book on digital video quality.

Christof Faller received the Ph.D. degree in computer and communication sciences from the Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, in 2004, and the M.Sc. (Ing.) degree in electrical engineering from ETH Zurich, Zurich, Switzerland, in 2000.

During his studies, he was an independent Consultant for Swiss Federal Labs, applying neural networks to process parameter optimization of sputtering processes, and spent one year at the Czech Technical University (CVUT), Prague, Czech Republic. In 2000, he became a Consultant and later Member of the Technical Staff for the Speech and Acoustics Research Department, Bell Laboratories, Lucent Technologies, where he focused on new techniques for audio coding applied to digital satellite radio broadcasting. At the Lucent spin-off, Agere Systems, he developed algorithms for parametric coding of multichannel audio signals, echo control, and other communications-related audio applications. Currently, he is with the Audiovisual Communications Laboratory at EPFL.